Share Analysis. Not Data.

PRANA-DATA

DB.1 White paper
Data-analyse zonder (leesbare) data

Project	PRANA-DATA
Project leader	Wessel Kraaij (TNO)
Work package	White paper
Deliverable number	DB.1
Authors	Paulien van Slingerland (TNO), Thymen Wabeke (TNO), Wessel Kraaij (TNO), Andre Dekker (MAASTRO), Marie José Bonthuis (UMCG)
Reviewers	Wessel Kraaij, Jessica Doorn
Date	31-03-2017
Version	1
Access Rights	Public
Status	Final

PRANA-DATA Partners:

Portavita, TNO, Radboud Universiteit Nijmegen, Maastricht UMC+, UMCG

COMMIT/ COMMIT is a public-private research community solving grand challenges in information and communication science shaping tomorrow's society

COMMIT/ A public- private research community

This publication was supported by the Dutch national program COMMIT.

Contents

Samenvatting

Innovatie op basis van (big) data: *Mag dit wel van wet- en regelgeving? Zijn de risico's op datalekken wel acceptabel? Accepteren mijn klanten dit wel? Accepteer ik dit zelf wel?*

Dit zijn veelgehoorde twijfels bij partijen die innovaties op basis van data overwegen. Deze twijfels zijn het gevolg van het feit dat bestaande oplossingen voor data-analyse over gecombineerde datasets vereisen dat één partij alle data verzamelt, interpreteert en analyseert. Dit whitepaper stelt innovatieve alternatieven voor, waarbij partijen geen gevoelige data hoeven uit te uitwisselen en toch tot nieuwe inzichten kunnen komen. Kortom: data-analyse zonder (leesbare) data.

1 Wat analyses over gecombineerde datasets in de weg staat

Wellicht volgt u de ontwikkelingen op het gebied van big data al langere tijd. En ziet u ook kansen in het gebruik van data voor meer efficiëntie, effectiviteit, nieuwe oplossingen, en uiteindelijk zelfs transformaties van bestaande markten.[1] En wilt u meer gebruik maken van data-analyses voor het verbeteren van besluitvorming, onderzoek, profilering van klanten, personalisering van producten en services, kortom: nieuwe verdienmodellen of verhoogde impact op basis van data.

> De zorgsector is een goed voorbeeld. Het is bekend dat de zorgkosten stijgen.[2] De Big Data Value Association rapporteert dat er naar schatting €330 miljard aan besparing mogelijk is in Europa, en dat juist big data technologie hiertoe uitgelezen kansen biedt in toepassingen als value-based care, precision medicine, lifestyle als medicijn, klinisch onderzoek op basis van *real word data*, etc.[3]

En toch komen voorgenoemde data-innovaties maar beperkt van de grond, vooral zodra er informatie van verschillende partijen gekoppeld dient te worden. Hiertoe zijn praktische oplossingen nodig om te zorgen voor de juiste mate van interoperabiliteit, datakwaliteit, afstemming van belangen, etc. Daarnaast dienen onderstaande hindernissen weggenomen te worden, die voortkomen uit de (privacy- of bedrijfs)gevoeligheid van de data. De rest van dit white paper is gericht op deze laatste categorie.

[1] http://www.idnext.eu/files/TNO-whitepaper--Big-data-in-small-steps.pdf
[2] http://www.zorgwijzer.nl/zorgverzekering-2018/kosten-zorgverzekering-stijgen
[3] Rapport: "Big Data Technologies in Healthcare", Big Data Value Association, TF7 Healthcare subgroup, 2016

1.1 Hindernis 1: Mag dit wel volgens wet- en regelgeving?

De wetgeving rondom de bescherming van persoonsgegevens is en wordt fors aangescherpt (zoals de nieuwe Europese Verordening voor bescherming van persoonsgegevens). De vertaling hiervan naar de praktijk is echter niet eenvoudig: wie mag wanneer welke persoonsgegevens delen en met wie? In hoeverre mogen gegevens hergebruikt worden? Hierdoor zoekt men in de praktijk op de tast naar de grenzen van doelbinding, proportionaliteit, etc. Dit is onder andere te zien bij de aanpak van kindermishandeling[4], fraudebestrijding[5], en bij de cross-sectorale aanpak voor personen met verward gedrag:

> "Privacywetgeving of in ieder geval de interpretatie daarvan staat professionals in de weg om relevante informatie over bijvoorbeeld het zorgverleden of strafdossier met elkaar te delen."[6]

1.2 Hindernis 2: Zijn de risico's op datalekken wel acceptabel?

Per 1 januari 2016 dienen organisaties een datalek te melden bij de Autoriteit Persoonsgegevens. In sommige gevallen dienen tevens de klanten die bij het lek betrokken zijn geïnformeerd te worden over het betreffende lek. De Autoriteit Persoonsgegevens kan organisaties een bestuurlijke boete opleggen van maximaal € 820.000 of 10% van de jaaromzet."[7] De nieuwe GDPR maakt nog hogere boetes mogelijk. Afgezien van deze sancties kan onrechtmatig gebruik van data via datalekken ook de reputatie en het vertrouwen van klanten aantasten. De verdeling van aansprakelijkheid in de keten maakt dergelijke risico's extra lastig te beheersen.

> "De Autoriteit Persoonsgegevens (AP) heeft [in 2016] bijna 5.500 meldingen van datalekken bij bedrijven en andere organisaties ontvangen. (…) Bijna één derde van de meldingen komt uit de gezondheidssector. Daarna volgen financiële dienstverlening (17 procent) en openbaar bestuur (15 procent). In september [2016] had al bijna de helft van de Nederlandse gemeenten een datalek gemeld."[8]

[4] http://www.huiselijkgeweld.nl/beleid/landelijk/overzicht-wet-en-regelgeving-uitwisseling-gegevens-bij-kindermishandeling
[5] http://www.wrr.nl/fileadmin/nl/publicaties/PDF-Working_Papers/WP_21_Big_Data_voor_fraudebestrijdingf.pdf
[6] https://vng.nl/publicaties/tussenrapportage-personen-met-verward-gedrag
[7] https://www.turien.nl/blog/kosten-van-een-datalek-zijn-enorm-7
[8] http://www.nu.nl/internet/4371533/bijna-5500-datalekken-gemeld-in-2016.html

1.3 Hindernis 3: Accepteren mijn klanten dit wel (reputatieschade)?

Zelfs als aan de wet wordt voldaan, kan het zijn dat klanten - of de maatschappij in brede zin - kritisch staan tegenover een initiatief. Dit kan voortkomen uit zorgen rond de bescherming van privacy (82% van de consumenten geeft aan “meer controle te willen over de gegevens die zij aan bedrijven verstrekken en de manier waarop bedrijven die gegevens gebruiken”[9]) of commerciële belangen (centralisatie van data is centralisatie van macht). Gepseudonimiseerde data mag gecombineerd worden ten behoeve van wetenschappelijk onderzoek, maar dit wordt steeds minder acceptabel vanwege het risico op re-identificatie (zo schat men dat 87% van de Amerikanen uniek te identificeren is aan de combinatie van postcode, geslacht en geboortedatum[10]). Negatieve aandacht in de media kan een initiatief compleet blokkeren en soms zelfs de reputatie schaden. Bekende voorbeelden hiervan zijn de proef van ING met het beschikbaar stellen van klantgegevens[11], het Britse CareData-programma met o.a. een deal tussen de National Health Service en Google’s deep mind[12], en het initiatief van Minister Schippers voor het koppelen van medische gegevens tussen zorgverleners en zorgverzekeraars[13].

> Nivel rapporteert over de weerstand van zorggebruikers over het hergebruik van hun medische gegevens: “26% zou bezwaar maken wanneer zijn gegevens zonder toestemming gebruikt werden als het om puur wetenschappelijk onderzoek ging, 33% als het om een farmaceutisch bedrijf ging, 43% als het om een ander bedrijf ging”, maar aan de andere kant: “een ruime meerderheid (75%) gaf aan niet zelf over die gegevens te hoeven beslissen als ze goed beveiligd zijn en uitsluitend voor [wetenschappelijk] onderzoek worden gebruikt”.[14]

1.4 Hindernis 4: Accepteer ik dit zelf wel?

Het delen van informatie met andere partijen brengt risico’s met zich mee, ondanks alle voordelen. Zo kan bedrijfsgevoelige informatie buiten de eigen controlezone komen; een partij kan soms aansprakelijk worden gesteld voor de gevolgen van het gebruik van zijn data; de waarde van de data kan verminderen en zo de concurrentiepositie verslechteren; derden kunnen kritiek op de bedrijfsvoering baseren op gedeelde informatie; zelfs de pricingstrategie kan mogelijk worden afgeleid:

> RTL Z meldt naar aanleiding van een handmatige data-analyse: "Jarenlang waren de marges van de opticiens veel te hoog. Die markt was zo gesloten dat niemand wist hoe duur een bril werkelijk was. De werkelijke prijs bleek een schijntje van de gevraagde prijzen."[15]

Al deze risico’s kunnen organisaties ervan weerhouden om iets met hun waardevolle data te doen ten behoeve van data-innovaties

[9] https://ddma.nl/wp-content/uploads/2016/06/DDMA_privacy-onderzoek-NL_def-4.pdf
[10] http://www.citeulike.org/user/burd/article/5822736
[11] http://www.nu.nl/politiek/3729359/geen-nieuwe-wetgeving-privacy-waarborgen.html?redirect=1
[12] http://www.zorgictzorgen.nl/big-data-analyse-zorg-is-afhankelijk-volatiel-publiek-vertrouwen/
[13] http://www.volkskrant.nl/wetenschap/kamer-niet-ingelicht-over-breder-delen-zorgdata~a4310578/
[14] http://postprint.nivel.nl/PPpp6202.pdf
[15] http://www.rtlnieuws.nl/economie/waarom-is-een-bril-vaak-zo-duur

2 Data-analyse zonder (leesbare) data?

De genoemde wettelijke beperkingen, risico's op datalakken en acceptatie door klanten en de eigen organisatie zijn een direct gevolg van de gebruikelijke aanpak voor data-analyse over gecombineerde datasets: namelijk dat één partij alle data verzamelt, combineert, en analyseert. Hiermee krijgt deze partij alle (gevoelige) informatie in handen, met alle risico's, machtsverschuivingen en weerstanden van dien. Wat als een *trusted third party* bijvoorbeeld wisselt van eigenaar of beleid?

Recent ontwikkelde technieken[16] bieden echter een alternatief: deze maken het mogelijk om informatie te analyseren die te allen tijde ofwel vercijferd (versleuteld) ofwel op de oorspronkelijke locatie blijft. Hiermee kan gevoelige informatie beschikbaar worden gesteld voor een specifiek doel, zonder deze gevoelige informatie zélf prijs te geven (vercijferde data heeft immers geen betekenis zonder ontcijfering). Hieronder staan enkele concepten op basis van deze technieken omschreven.

2.1 Concept 1: Clusteren van vercijferde data (zonder deze op enig moment te ontcijferen)

Waarom? Het verrijken en vergelijken van gebruikersprofielen op basis van gegevens van andere partijen leidt over het algemeen tot betere en nieuwe inzichten. Door vergelijkbare patiëntendossiers naast elkaar te leggen kunnen patiënten in zekere zin elkáár helpen genezen. Doorgaans kan dit echter niet zonder het ongewenste neveneffect dat de beheerder van het profiel toegang krijgt tot de aanvullende persoonsgegevens.

Dit concept maakt een clustering van personen (of andere entiteiten) op basis van vercijferde gegevens van meerdere partijen over deze personen. Door deze nieuwe vorm van vercijfering (homomorfe encryptie) blijven de gegevens te allen tijde beschermd. De relatief hoge benodigde opslagcapaciteit en rekenkracht vereisen wel dat het datavraagstuk zodanig wordt geformuleerd, dat er relatief weinig gegevens uitgewisseld hoeven te worden. Ter indicatie: clustering van 100.000 personen op basis van 12 eigenschappen duurt ongeveer anderhalf uur.[17]

Bijvoorbeeld: In het project PRANA-DATA is een proof-of-principle ontwikkeld voor het analyseren van jeugdgroei door vercijferde gegevens van verschillende consultatiebureau's te combineren. Specifiek kan dit proof-of-principle de verwachte groeicurve van jonge baby's voorspellen. Daarnaast zijn andere statistieken te verkrijgen, zoals: *X% van de vergelijkbare kinderen had last van een groeiachterstand. Hiervan had Y% baat bij aanbeveling A*. De groeivoorspelling is gebaseerd op historische data van vergelijkbare kinderen. Deze data is homomorf vercijferd voordat het werd gedeeld en dus niet zichtbaar voor de analyserende partij. Op een vergelijkbare manier zouden ook inzichten in de effecten en complicaties van bijvoorbeeld medicatie bepaald kunnen worden op basis van vercijferde gegevens bij meerdere zorgverleners.

[16] Het gaat hier om technieken zoals *secure multi-party computation*, *homomorfe encryptie* en *online machine-learning*. De mogelijkheden van blockchain technologie worden hier buiten beschouwing gelaten.
[17] http://jis.eurasipjournals.springeropen.com/articles/10.1186/1687-417X-2013-4

2.2 Concept 2: Rekenen met gedistribueerde data (zonder dat deze van zijn originele plek komt)

Waarom? Het combineren van data van meerdere organisaties levert over het algemeen betere wiskundige modellen voor het maken van voorspellingen. Door gegevens van patiënten uit verschillende ziekenhuizen te combineren kan bijvoorbeeld een beter model getraind worden om de levensverwachting van een terminale patiënt te kunnen schatten. Normaal gesproken kan dit echter niet zonder dat één partij alle data-input in handen krijgt.

Dit concept brengt de algoritmen naar de data in plaats van andersom. Hierbij wordt een model getraind op basis van de data van één partij, om dit model vervolgens te verbeteren op basis van data van een tweede partij, etc. Dit iteratieve proces wordt herhaald totdat de juiste kwaliteit is bereikt. De onderliggende gegevens blijven ondertussen te allen tijde op hun originele locatie bij de beherende instantie. Deze aanpak vereist wel dat de gecombineerde databronnen gelijksoortige informatie bevatten (over verschillende gebruikers). Het is dus niet ontworpen voor het combineren van verschillende soorten informatie over één enkele gebruiker.

Bijvoorbeeld: De Personal Health Train[18] ontwikkelt deze aanpak om medische modellen te trainen op basis van historische data van meerdere ziekenhuizen. De data verlaten het ziekenhuis niet; de medische modellen wel.

2.3 Concept 3: Vergelijken van vercijferde data (zonder deze op enig moment te ontcijferen)

Waarom? Om te valideren of bepaalde informatie correct is, kan een vergelijking met een externe betrouwbare bron waardevol zijn. Voor de politie kan het bijvoorbeeld waardevol zijn om te weten of een persoon met verward gedrag op straat in een bepaalde database van de GGZ voorkomt.[19] Normaal gesproken kan dit echter niet zonder dat minstens één van de twee partijen aanvullende informatie prijsgeeft (het feit dat de vragende partij kennelijk ‘interesse’ heeft in de betreffende persoon kan gevoelige informatie op zich zijn).

Dit concept, Trust Tester[20], voert de beoogde vergelijkingen uit met vercijferde data (zonder deze op enig moment te ontcijferen). Hierdoor leert geen enkele betrokken partij iets dat hij nog niet wist, en krijgt de ontvangende partij alleen de zekerheid of een uitspraak correct is of niet. Deze aanpak is niet ontworpen voor complexe rekenmodellen, maar voor vragen van de vorm: is X groter dan (of gelijk aan) Y?

Bijvoorbeeld: TrustTester kan gebruikt worden voor het automatiseren van hypotheekaanvragen. Hierbij wil de bank kunnen valideren of het door de klant opgegeven (minimum) inkomen inderdaad klopt door deze te vergelijken met inkomensgegevens bij een publieke instantie. Hierbij leert de bank niet het daadwerkelijke inkomen en de publieke instantie leert niets over de hypotheekaanvraag. Toepassingen in andere domeinen zijn ook goed mogelijk.

[18] http://www.dtls.nl/fair-data/personal-health-train/
[19] https://vng.nl/files/vng/publicaties/2016/20160705-avp-2e_tussenrapportage-def.pdf
[20] http://pimn-public.sharepoint.com/Documentatie/2016-07-00-idm-Rijksoverheid-Essay%20Persoonlijk%20Data%20Management%202016.pdf

2.4 Concept 4: Deelnemen aan een veiling waarbij elk bod beschermd blijft (door encryptieprotocollen)

Waarom? Via een gesloten veiling, waarbij elk bod vertrouwelijk blijft, kan een eerlijke marktprijs voor een product worden bepaald op basis van vraag en aanbod. Een dergelijke aanpak is mogelijk een waardevol element in de huidige ontwikkeling van value-based care, waarbij de effectiviteit en kosten van een behandelpad vooraf, tijdens en na de behandeling geëvalueerd worden.[21] Normaal gesproken is een gesloten veiling echter niet mogelijk zonder dat een dienstverlenende trusted third party inzicht krijgt in de economische positie van de bieders.

Dit concept heeft een biedsysteem op basis van encryptieprotocollen. Hiermee kan vraag en aanbod op elkaar aangepast worden, zonder dat een derde partij ieder bod te weten komt.

Bijvoorbeeld: In Denemarken hebben enkele duizenden boeren deelgenomen aan een suikerbietenveiling op basis van deze aanpak.[22]

3 De voordelen van data-analyse zonder (leesbare) data

> Michael van den Berg (verbonden aan het Rijksinstituut voor Volksgezondheid en Milieu): "Zonder koppeling van data geen goede zorg. Door eenzijdige aandacht voor privacy zien we de noodzaak van analyse en koppeling van zorgdata niet."[23]

De uitdaging is dus om én aandacht te kunnen hebben voor privacy én koppeling van data te realiseren. De hiervoor besproken concepten zijn gericht op beide wensen. Met dergelijke concepten kunt u:

1. samen met partners en klanten tot nieuwe analyses van gecombineerde datasets komen (zonder privacy- of bedrijfsgevoelige informatie uit te hoeven wisselen),
2. uw eigen data beschermen,
3. vertrouwen van klanten behouden,
4. concurreren op bescherming van privacy,
5. negatieve media-aandacht vermijden,
6. centralisatie van data vermijden (kennis is macht),
7. security risico's van grootschalige databases vermijden,
8. participatie aan data-analyses over gecombineerde datasets per toepassing afwegen en besluiten (controle).

Op deze wijze draagt de technologie bij aan het wegnemen van de eerder genoemde hindernissen omtrent compliance, datalekken, en acceptatie. Contracten, convenanten en procedures kunnen immers worden gebroken. Door óók op fundamentele oplossingen in te zetten, staat gevoelige data 'dubbel op slot'. Dat geeft vertrouwen.

[21] Rapport: "Big Data Technologies in Healthcare", Big Data Value Association, TF7 Healthcare subgroup, 2016
[22] https://www.cs.purdue.edu/homes/aliaga/cs197-10/papers/bogetoft.pdf
[23] http://www.volkskrant.nl/opinie/zonder-koppeling-van-data-geen-goede-zorg~a4321107/

4 Over PRANA-DATA

Het project PRANA-DATA (Privacy Respecting ANAlysis of distributed patient health DATA) heeft privacy vriendelijke manieren ontwikkeld om data te delen: de toegang tot persoonlijke gegevens wordt vermeden dan wel tot een minimum beperkt door een aantal maatregelen:
- het toepassing van decentrale privacy respecterende analytse modellen (gedistribueerd);
- het combineren van data van verschillende gedistribueerde bronnen zonder de data zelf vrij te geven;
- het toepassen van nieuwe geavanceerde cryptografische technieken zoals secure multi-party computation, homomorphic encryption;
- het uitvoeren van de algoritmen bij de data zelf
- het op een slimme manier combineren van modellen die onstaan zijn uit analyses van meerdere bronnen.

> "The aim of PRANA-DATA is to explore the potential approaches for collecting, storing, combining and analyzing health related distributed personal data in a unified secure and privacy respecting system architecture that supports the interests of stakeholders of different domains, i.e. patients, LSH and medical researchers, health professionals, health policy makers and companies (i.e. pharma, food, insurance)."

Voor meer informatie, zie www.PRANA-DATA.nl.

5 Contact

Herkent u de geschetste uitdagingen rondom het delen van data-analyses? Wilt u de mogelijke oplossingen voor uw toepassing nader verkennen? Neem gerust contact met ons op via:

Wessel.Kraaij@tno.nl of Jessica.Doorn@tno.nl